Master DS: Applied Statistics

contact: jean-marc.freyermuth@univ-amu.fr

2024-2025

Summary

- Aims and guidelines of the course
- Some advises for statistical consulting
- stages of a statistical investigation
- Data science; Big Data paradises and paradoxes
- Structure in complex data

Section 1

Aims and guidelines

Aims of the course

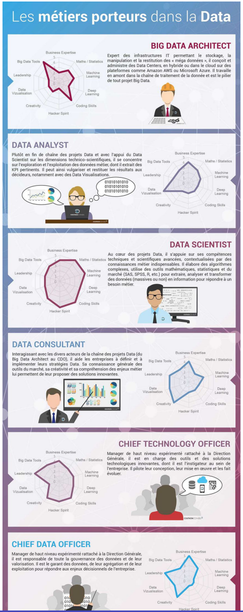
To prepare yourself to the real life of a data scientist. Alternative lecture title: *Statistical consulting*.

classical pedagogical approach in statistics: learn about a family of statistical methods and apply them to some real data examples.

A. Ehrenberg: "... the kind of examples of statistical analysis that tends to be considered in professional discussions are so grossly over-simplified as to make a pretentious mockery of real-life situations and statistical consultancy."

- **in real life**, data set (± well organized and treated) + questions (± well-formulated) → which method(s) to use to provide appropriate and accurate answers ?
    - wide range of possible methods to be used,
    - statistical methodology, statistical analysis pipeline may involve several "steps", think about the overall analysis pipeline is a must,
    - **answer to the client questions** this is conditioning the most appropriate choice for the client.
- the role of the statistician starts far "before" the statistical analysis (planing of experiment...).
- Essential aspect of statistical consulting: **COMMUNICATION**

Jobs in data science



Les métiers porteurs dans la Data

BIG DATA ARCHITECT

Expert des infrastructures IT permettant le stockage, la manipulation et la restitution des « méga données », il conçoit et administre des Data Centers, en hybride ou dans le cloud sur des plateformes comme Amazon AWS ou Microsoft Azure. Il travaille en amont dans la chaîne de traitement de la donnée et est le pilier de tout projet Big Data.

DATA ANALYST

Plutôt en fin de chaîne des projets Data et avec l'appui du Data Scientist sur les dimensions technico-scientifiques, il se concentre sur l'exploration et l'exploitation des données métier, dont il extrait des KPI pertinents. Il peut ainsi vulgariser et restituer les résultats aux décideurs, notamment avec des Data Visualisations.

DATA SCIENTIST

Au cœur des projets Data, il s'appuie sur ses compétences techniques et scientifiques avancées, contextualisées par des connaissances métier indispensables. Il élabore des algorithmes complexes, utilise des outils mathématiques, statistiques et du marché (SAS, SPSS, R, etc.) pour extraire, analyser et transformer des données (massives ou non) en information pour répondre à un besoin métier.

DATA CONSULTANT

Interagissant avec les divers acteurs de la chaîne des projets Data (du Big Data Architect au CDO), il aide les entreprises à définir et à implémenter leurs stratégies Data. Sa connaissance générale des outils du marché, sa créativité et sa compréhension des enjeux métier lui permettent de leur proposer des solutions innovantes.

CHIEF TECHNOLOGY OFFICER

Manager de haut niveau expérimenté rattaché à la Direction Générale, il est en charge des outils et des solutions technologiques innovantes, dont il est l'instigateur au sein de l'entreprise. Il pilote leur conception, leur mise en œuvre et les fait évoluer.

CHIEF DATA OFFICER

Manager de haut niveau expérimenté rattaché à la Direction Générale, il est responsable de toute la gouvernance des données et de leur valorisation. Il est le garant des données, de leur agrégation et de leur exploitation pour répondre aux enjeux décisionnels de l'entreprise.

Temptative schedule

Content (S1)

1. On some ‘principles’ of applied statistics; modern challenges in statistics: a biased point of view;
2. Real Case study 1. *A stock management problem.* 3-4) An historical perspective on two sample tests: from the Behrens-Fisher problem to modern applications in brain imaging. Focus on some general statistical modelling principles, paradoxes, modern research topic. . .

Content (S2)

TBA (a review on PCA, topologically augmented ML, an example of data analysis pipeline)

Lecture guidelines

focus on the “second order” messages; active involvement during the lecture.

- Grading: 2 group projects
- Advice
  - how to write this report: → be client oriented.
- on Ametice
  - some notes, side readings, slides,
  - some useful links,
  - some code used during the class,
  - **the data of the consulting project**.

Grading of PART II

consulting projects ***alike a Data Science consulting game****: the best team is the one that would get the contract with our client! Projects will be ranked to be graded.*

written report

- Communication
  - quality of the document (writing, structure, visual, graphs)
  - understanding of the client problem
  - reproducibility
- Statistical approach to the problem
  - choice of proper methodology to answer the client questions (descriptive statistics, modeling, learning. . .)
  - quality of the statistical treatment (model fitting, interpretation of the results)

oral defense (10-15 min)

- *Ethos* and *Logos* matters. . .

Section 2

Some advises for statistical consulting

How to prepare oneself for statistical consulting ?

personal aspect of the consultant

- Bisgaard and Bisgaard (2005) distinguish three types of consultants:
    - those doing low level jobs,
    - the expert,
    - the coach, the colleague.
- **no ego, we go**, being client-oriented, the will to answer to the client questions, sounds obvious, but. . .
- **self-denial** “you are a statistician that will take you just a few minutes”
- **social psychology** be attentive to the body language of the client, guess the “unsaid”. Its motivations could be very different from those expressed verbally (trying to cover oneself up to sthg, e.g., create internal conflict in his own company. . .).

How to prepare one self for statistical consulting ?

personal aspect of a consultant

- **ability for active listening**:
  - **Third king error in consulting**, Kimball (1957): *"the error committed by giving the right answer to the wrong problem"*.
  - help the clients ? formulate the questions in terms that can be quantified.
- **be learned** wide knowledge and true understanding of statistical methods but also more generally *scientifically knowledgeable*.
- **handyman** ability to adapt statistical methods to new environments.

How to prepare one self for statistical consulting ?

Are they general principles that can be formulated ?

Yes, but beware of numerous application fields that have specific connection to some methodologies.

D.R. Cox, in *Principles of applied statistics:*

"I am broadly very sympathetic to the idea that success often hinges more on scientific sense than on technical mastery of complex methods"

A detour on reproducibility (1)

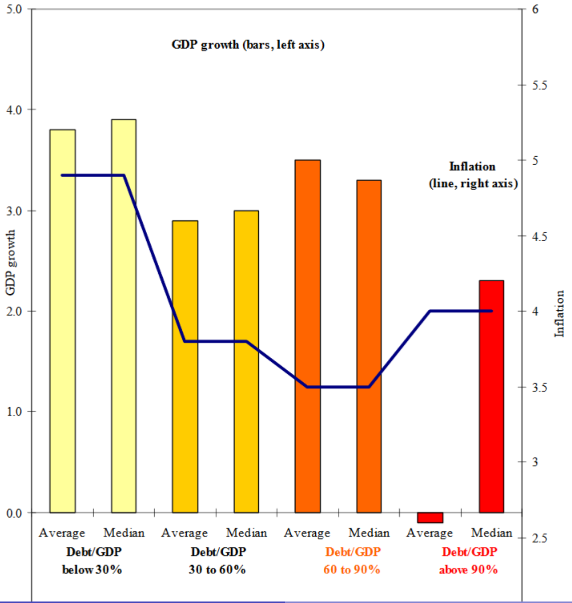
Ref: David Louapre - Youtube channel ‘ScienceEtonnante’

Reinhart, C. M., & Rogoff, K. S. (2010). **Growth in a Time of Debt**. American economic review, 100(2), 573-78.

Dataset: observation of 3700 annual observations of economic growth and inflation at different levels of government and external debt.

Main finding: there is a threshold at 90 percent of GDP above which the average growth falls considerably !

A detour on reproducibility (2)



A detour on reproducibility (3)

https://www.bbc.com/news/magazine-22223190

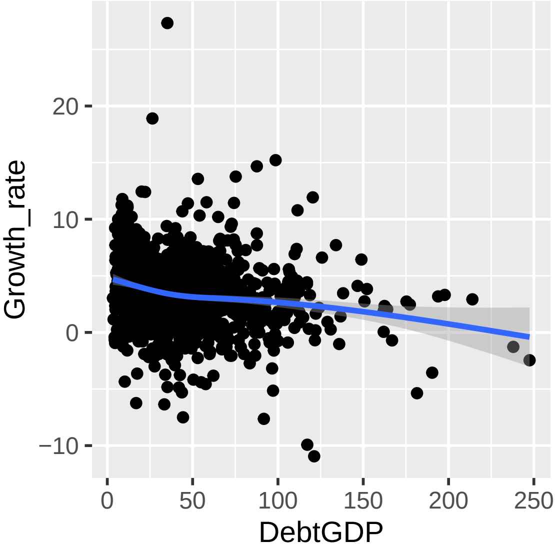
Herndon, T., Ash, M., & Pollin, R. (2014). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. Cambridge journal of economics, 38(2), 257-279.

Donoho, D. (2010). An invitation to reproducible computational research. 11, 3, 385-388.

Is a concern also for consulting too. Nowadays many tools help:

- Rmarkdown / notebook documents
- github
- docker. . .

A detour on reproducibility (4)



Section 3

5 stages of a statistical investigation

Stages of statistical investigation (1)

- 1-formulate the objectives, the questions
- 2-design of experiment
- 3-measures / data harvesting
- 4-data analysis
- 5-interpretation
- (6-presentation of results)

Academic training generally concentrates on (2)-4-(5)...

Stages of a statistical investigation (2)

Within these different stages, we wonder of about existing interactions between **the domain of application and the statistical aspects**. The latter are quite clear for points 1-2-5 but complicated for statistical analysis (whether it concerns statistical modeling, learning, or just descriptive statistics)

if statistical methods are closely related to a specific field of application, then what is the place of Statistics as a Science ?

Stage 1: formulate the objectives

the role of statistics: what **can** and **cannot** be highlighted from a dataset.

There is a need to distinguish:

- inferential, i.e., contribute to a better understanding of a phenomenon → by doing inference on the *Data Generating Process*, causal study... often there is a series of hypothesis to formulate (research questions) to which we aim at answering based on statistical results.
- help with decision → bayesian modeling.

Stage 2: experimental design (1)

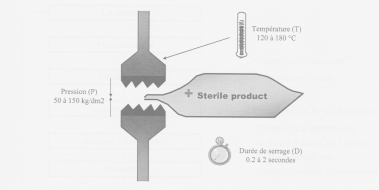
- a problem in the experimental design may affect all the statistical analysis (identifiability, estimability)
- **objectives:** find the structure of an experiment, which data to collect to reach the objectives. I.e., control the variability, the power... while controling the **costs** (impossible to capture of sources of variability and test each combination)

Your client is **the expert** of his own data and of his problem, the statistician is an expert in uncertainty quantification. Be really careful on the plausible **causes of variability** and if they can be controlled or not.

Stage 2: experimental design (2)

Thermo-closure of bags containing sterilized materials.

Manufacturing process depends on 3 parameters(definition of the experimental domain)



But : rechercher des conditions de fabrication qui assurent une résistance à la soudure entre 75 et 85 (kg/dm^2)

Stage 2: experimental design (3)

Objectives:

- find the optimal manufacturing settings to obtain a pre-specified welding resistance (measured in Kg/dm^2)

Questions:

- do we need replicates, if yes, how many ? How many points in the parameter space to achieve a given accuracy of prediction under costs constraints.
- is there an order to follow in the experiences ?

Stage 3: collecting the data (1)

The client is the expert of his own data, the statistician is expert in dealing with uncertainty and about **cascading risks and uncertainty**.

> *"What is the meaning of the data finally acquired and that we have to analyze"*

- **relevance:** capture the essence of DGP. . . towards **interpretability**
- **precision:** repeated measurements by the same or different operators would give similar results ? (comparing measurement equipments)
- **efficiency** costs and quality/quantity of data
- **collateral effects** of taking measurements; **Heisenberg uncertainty principle**, is the observation of a system having an impact on the measures taken on this system ?
- storage/exploiting the data

Stage 3: collecting the data (2)

typology of measurement scales

- continuous
- categorical
- interval. . .

typology by nature

- response variables
- explanatory variables
- _primairy
- _potentially causal
- _contextual variables
- _non specific (e.g multi-centre data)

typology by "physical nature":

- additive or not

Stage 4 data analysis (1)

Several stages:

- **data auditing:** transform the raw data to data that can be exploited
- **clearness:** the link between data and interpretation should be as clear as possible, i.e., Occam's Razor principle *use methods the simplest and the most appropriate* **EVEN IF... drifts exist**.
- **fragmentation:** clearness can be improved by dividing the data onto several subgroups for separate analyses
- **definition and model choice:** a model has to capture the essence of the GDP, i.e., of deterministics and stochastic components.
  - data driven
  - model based (using models developped within the field of application): choice of the parametrization and critic (is there a parameter answering the question ? does it have an interpretation that make sense within the application field).

Stage 4: data analysis (2)

- graphical modelling of **dependencies** (DAG)
- **randomness** multilevel, hierarchical
- **model choice** empirical, example of linear regression, choice for interpreting the variables, predictive power, explanatory power (R2, prediction, PLS, PCR)
- Evaluating the accuracy of the conclusions, variability estimation, precision vs variability
- Reformulate the objectives: check the ability to interpret the different patterns that emerge from formal data analysis.

Stage 5: interpreting the results

The utmost a model shows a strong *field-method* link, the utmost the interpretation will be performed with ease.

Stage of formalizing the results

- translate established facts to benefits, example of BVA (first real case study), build a prediction device directly usable by technical staff
- this stage consists in a conscious stage of transferring statistical results to the client language. This implies

```
+ identify the consequence of discoveries for the client,
+ understand the problem in its context
+ efficient communication with the client.
```

presentation of the results

- Show that it was useful and do not discard from their main problem.
- The concept and models of statistical nature have very often **not to** be presented !
  - use graphics, animations within a sober style.

Communication (1)

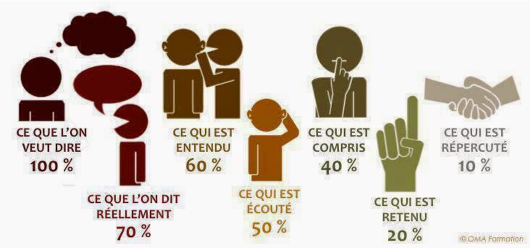


Figure 1: 1 picture worth 1000 words

Communication (2)

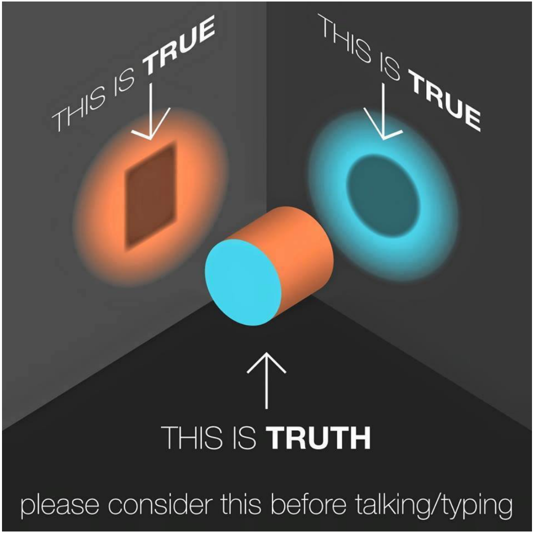


Figure 2: 1 picture worth 1000 words

Section 4

Data science - Big Data paradises and paradoxes

Science of data Science

Statistician vs data scientist ?

Is DS a real science:

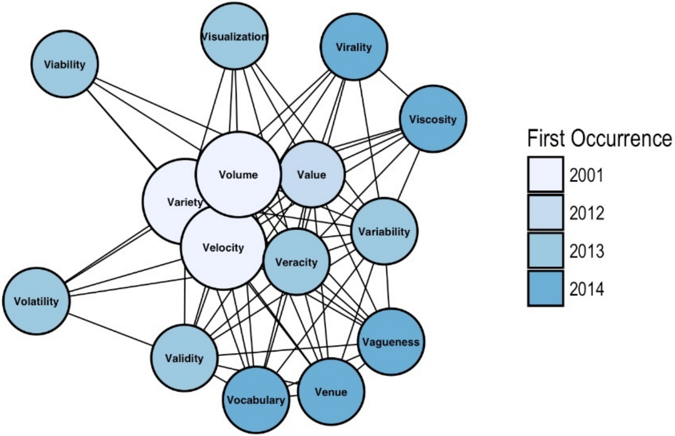
- intellectual content
- organized in a comprehensible form
- use of the real life observations as a validity test (falsifiability)

Lectures:

- Tukey 1964: The future of data analysis
- Breiman 2001: Two cultures
  - contrast a world where 98% of approaches are done using modelling against 2% with learning (very different now !)
  - "if our goal as a field is to solve problems, then we need to move from exclusive dependence on models"
- Donoho 2017: 50 years of data science.

definition(s) of big data the 4V's and beyond

- definition of Big Data: non unique, field-specific... **Bell (2015): I shall not today further attempt to define Big Data, but I know it when I see it".**
- *Shafer, T (2017), "The 42 V's of Big Data and Data Science". https://www.elderresearch.com/blog/42-v-of-big-data



definition(s) of big data the 4V's and beyond

Volume

- Need of special infrastructure (Hadoop, clusters. . .)
- impossible to access the whole dataset on a given processor ⇒ many methods will fail because they are unable to be used on separate pieces of the data **scalability**

Variety

- related to an increase in complexity: audio, images, network
- often high-dimensional data ($n >> p$), e.g. genomic data.
- challenge classical understanding of asymptotic properties of statistical methods

definition(s) of big data the 4V's and beyond

Veracity

- multiple data sets put together without being collecting with the current research question in mind, known as "convenience samples"
- classical ideas of experimental design, survey sampling are lost in favor of availability of large amount of data ⇒ issues alike **sampling bias** need not to be ignored !

Velocity

- here big means, data arriving at high velocity; e.g., high frequency financial data, particle physic experiments. Need **online** data analysis.

definition(s) of big data the 4V's and beyond

- Big evolves with computational resources
- Finally, may be it is not so much about *the data amount* but
  - high-dimensional
  - sequential
  - sparse
  - complex
- important aspect pointed out by Xiao Li Meng: *Big Data often involves a conglomeration of datasets-most of which are not collected for inference purpose.*

Some examples of massive data

- genomic data: half millions of micro networks are now publicly available, each of these contains tens of thousands of values of molecular expressions
- fMRI: 3D videos of images made of tens of thousands of voxels
- astrophysical data: LSST (Large Synopic Survey Telescope), records 30 To of data each nights
- text data
- financial data
- video surveillance. . .

Big data and learning

Recent definitions of Big-Data associate them to machine learning algorithms that are able to deal with them.

shared enthusiasm

“Cette intelligence artificielle dont le champ d’application s’étend désormais du diagnostic médical à la voiture autonome et à la distributioon d’énergie, pour ne citer que quelques exemples, est aujourd’hui au coeur de préoccupations industrielles et politiques majeures” (C. Villani, 2018).

and fear

- machine learning algorithms are more and more discussed and disputed
- projects using big data and their algorithms are mainly controlled by private companies and can be under hidden control on purpose
- many ethical questions have been raised by the use of such algorithms
  - confidentiality
  - transparency
  - discriminatory bias. . .

Big data and learning

Cathy O'Neil (2016) : un algorithme n'est en réalité qu'une “opinion intégrée aux programmes”

Le rapport de la commission Villani appelle à “ouvrir les boites noires” de l'IA et considère qu'il s'agit d'un enjeu démocratique.

Un développement inéluctable qu'il faut savoir démystifier et encadrer P. Besse et al., 2019 : “Aborder sérieusement ces questions nécessite à la fois de sérieuses compétences techniques, afin de comprendre finement le fonctionnement des algorithmes et de garder un regard critique sur le discours qui les entoure, et une expertise juridique, sociétale ou sociologique, voire politique ou philosophique.”

HG. Wells : “Le jugement statitisque sera un jour aussi nécessaire à l'exercice de base des fonctions du citoyen que la capacité de lire et d'écrire.”

Statistical paradises and paradoxes in Big Data

Xiao Li Meng (2018) Statistical paradises and paradoxes in Big Data... The annals of applied statistics, 12(2), 685-726

Answers the questions:

- is an 80% non random sample better than a 5% random one?
- Should you trust more a survey with 60% of response rate or non probabilistic data set covering 80% of the population ?

Statistical paradises and paradoxes in Big Data

Consider a finite population of size N. Let $X_j,\ 1 \leq j \leq N$ be observations on the j-th individual (e.g., vote for Trump, Yes = 1 or No =0; size in cm . . .)

Population quantity can be expressed as a population average $\bar{G}_N$ of $\{G_j \equiv G(X_j); 1, \dots, N\}$ (choosing an appropriate function G)

$\bar{G}_n$ is a sample computed on a subset I_n of size n of $\{1, \dots, N\}$

$$\bar{G}_n = \frac{1}{n} \sum_{j \in I_n} G_j = \frac{\sum_{j=1}^{N} R_j G_j}{\sum_{j=1}^{N} R_j}$$

where

$$\begin{cases} R_j = 1 \text{ for } j \in I_n \\ R_j = 0 \text{ else.} \end{cases}$$

Hope of big data, since n is large everything should be fine ! **BUT** often omit data quality.

Statistical paradises and paradoxes in Big Data

We can show that

$$\bar{G}_n - \bar{G}_N = \underbrace{\rho_{R,G}}_{Data\ \ quality} \times \underbrace{\sqrt{\frac{1-f}{f}}}_{Data\ \ quantity} \times \underbrace{\sigma_G}_{Pb\ \ difficulty} ,$$

where, $f = n/N$ is the sampling rate. and, under the assumption of no response bias, under any R-mechanism.

Statistical paradises and paradoxes in Big Data

$$
\begin{aligned}
MSE_R(\bar{G}_n) &= \mathbb{E}_R[(\bar{G}_n - \bar{G}_N)^2] \\
&= \mathbb{E}_R[\rho^2_{R,G}] \times \left(\frac{1-f}{f}\right) \times \sigma^2_G \\
&:= D_i \times D_O \times D_U.
\end{aligned}
$$

- D_I Defect Index
- D_O Dropout odds
- D_U Degree of Uncertainty

Statistical paradises and paradoxes in Big Data

- paradox **the bigger the data, the bigger we fool ourselves**
- e.g., 2016 US presidential election $\rho_{R,X}$ very small, the sample proportion of voter for Trump computed from $n \approx 2.3e+06$ self reported voters has the same mean squared error than a simple random sample of 400 (effective sample size)...

To keep in mind:

- data quality is what matter the most, not quantity,
- consider the effective sample size, not the absolute sample size,
- don't ignore seemingly tiny probabilistic datasets when combining data sources,
- ensure that the recording mechanism is the less correlated as possible with the recorded value/variable,
- probabilistic sampling powerful tool to achieve this goal.

Section 5

Structure in complex data

The curse of dimensionality

Consider n observations in the following nonparametric regression problem:

$$Y_i = f(x_{i,1}, \ldots, x_{i,d}) + \epsilon_i,\ 1 \leq i \leq n,\ \epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1),$$

$f \in \mathcal{F} := \left\{ f : [0,1]^d \to \mathbb{R},\ \text{Lipschitz} \right\}$, then, for any estimator $\hat{f}_n$ of f:

$$\sup_{f \in \mathcal{F}} \mathbb{E} \left\| \hat{f}_n - f \right\|^2 \geq C n^{-\frac{2}{2+d}},$$

where C does not depend on d.

Curse of dimensionality term introduced by *Bellman, R. (1961). Adaptive Control Process: a guided tour. Princeton University press.* In the context of optimization of functions on a continuous domain, he contrasted: exhaustive enumeration strategies and dynamic programming algorithms.

The curse of dimensionality ?

Scott (1992, Chapter 2) : *"the underlying structure of d dimensional data is almost always of dimension lower than d"*.

effective vs observed dimension

- advantage of having some underlying structure in the data, this enables "*to avoid*" the curse of dimensionality, to improve interpretability. . .
- in the previous model, adaptive estimation of f w.r.t regularity and possible underlying structure
- modelling this structure:
  - Sobol decomposition $f \in L_2\,[0,1]^d$ (this decomposition exists and is unique if and only if the different **functional effects** are orthogonal)

$$f(x_1, \ldots, x_d) = \sum_{u=1}^{d} f_u\,(x_u) + \sum_{u<v} f_{uv}\,(x_u, x_v) + \ldots + f_{1,\ldots,d}\,(x_1, \ldots, x_d)\,.$$

Marginal components (main effects)

$$f_u : \;[0,1) \longrightarrow \mathbb{R},\;(1 \leq u \leq d).$$

The curse of dimensionality ?

We define the **atomic dimension** δ of f which reflects the maximal degree of intereaction between the d variables of f.

Additive structure:

$$f(\underline{x}) = \sum_{u=1}^{d} f_u(x_u) \rightarrow \delta(f) = 1.$$

Sructure with interaction of level $m < d$:

$$f(\underline{x}) = \sum_{u_1 < \cdots < u_m} f_{u_1,\cdots,u_m}(x_{u_1}, \cdots, x_{u_m}) \rightarrow \delta(f) = m.$$

- *compound functional models*

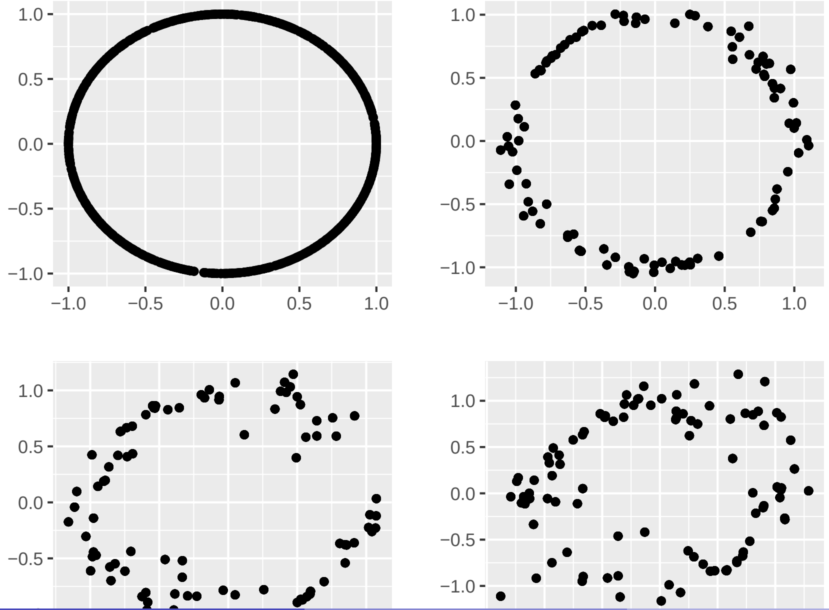
Arnak S. Dalalyan, Yurii Ingster, Alexandre B. Tsybakov (2014). Statistical inference in compound functional models. Probab. Theory Related Fields, 158(3), pp. 512-532.

Manifold learning

```
## wgl
##   1
```

```
## wgl
##   2
```

Topological data analysis: data have shape



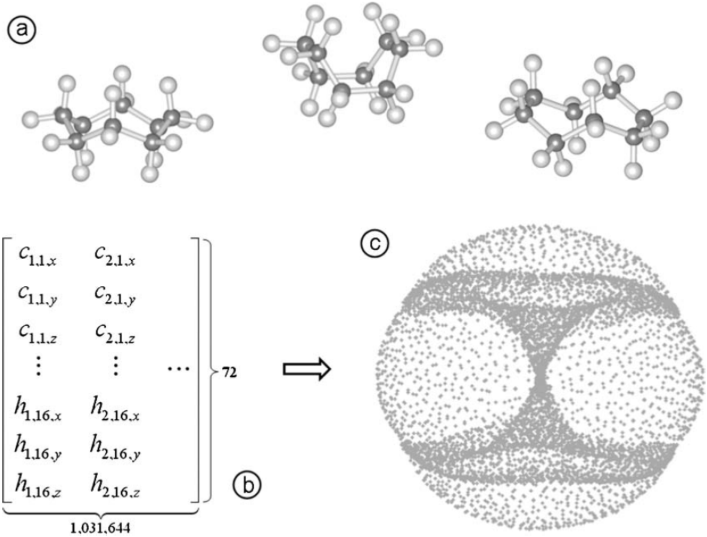
Topological data analysis

Martin, S et al (2010). Topology of cyclo-octane energy landscape. The journal of chemical physics, 132.

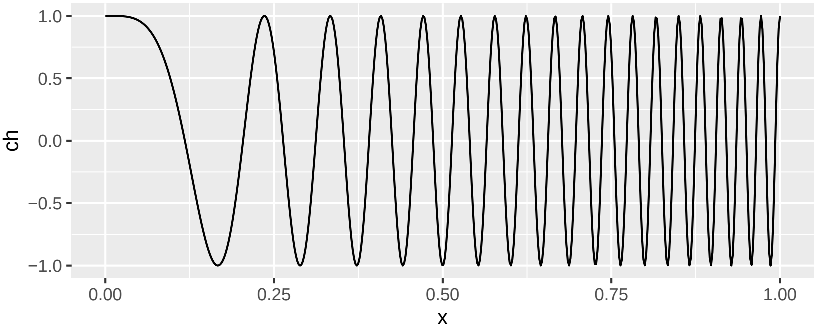
Aim: Understand the energy landscape of cyclo-octane ($C_8 H_1 6$). The energy landscape describe all possible conformations of a molecule.

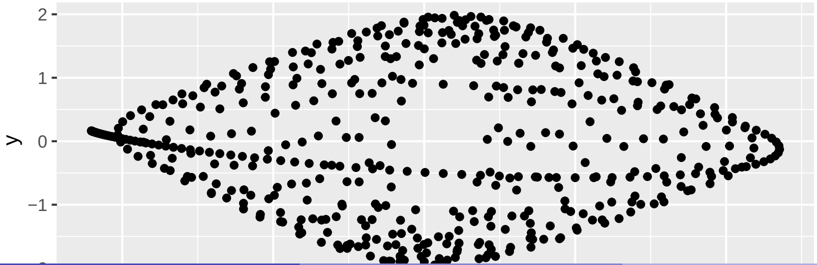
1. generate large data set $> 10^6$ cyclo-octane conformation
2. each conformation is place in cartesian space through 3D position coordinates of the atoms in the molecule
3. position concatenated in a vector of $\mathbb{R}^{72}$.

Topological data analysis


a
$c_{1,1,x}$ $c_{2,1,x}$
$c_{1,1,y}$ $c_{2,1,y}$
$c_{1,1,z}$ $c_{2,1,z}$
$\vdots$ $\vdots$ $\cdots$
$h_{1,16,x}$ $h_{2,16,x}$
$h_{1,16,y}$ $h_{2,16,y}$
$h_{1,16,z}$ $h_{2,16,z}$
72
1,031,644
b
c

Topological data analysis for time series data





Topological data analysis

- geometry:
  - quantitative (centrality, dispersion)
  - algebraic relations (rigid transformation, e.g.)
- topology:
  - qualitative (dimension, number of connected components, betti numbers);
  - algebraic invariant; homeomorphic, homotopic equivalence class

TDA:

- recent field using tools from applied algebraic topology and computational geometry
- aim: extract topological features from data represented as point of clouds in euclidean space